Naval Research Laboratory

Washington, DC 20375-5320

NRL/FR/5550--05-10,126

Hiding Information Under Speech

GEORGE S. KANG
THOMAS M. MORAN
DAVID A. HEIDE

Transmission Technology Branch
Information Technology Division

December 12, 2005

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
12-12-2005	Formal Report	October 1, 2004 — September 30, 2005

4. TITLE AND SUBTITLE

Hiding Information Under Speech

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
33904N, 61553N

6. AUTHOR(S)

George S. Kang, Thomas M. Moran, and David A. Heide

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
IT-235-009

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/FR/5550--05-10,126

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375

10. SPONSOR / MONITOR'S ACRONYM(S)

NRL

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Jerry Weisman of Santa Clara, CA, who has been coaching CEOs for many years on how to sell themselves, wrote about making successful presentations. Weisman said, "simple but accurate charts are essential to making a presentation more effective." Unfortunately, the secure phone has been no more functional than a telephone, i.e., a device capable of transporting only sounds (the human voice and surrounding sounds). We think that a future secure phone should move beyond the 20th century abilities and be capable of transmitting essential visual aids (such as key words, key phrases, and simple graphics, or hand-scribbled notes) to complement voice communication. We are implementing such a phone, operating in a real-time demonstrable prototype, in which a Variable Data Rate (VDR) voice encoder, which is currently being developed by SPAWAR PMW-160 as a future Navy secure phone for future IP applications, is the core voice encoder. With the technology described in this report, two tactical commanders can coordinate a tactical assault plan not only verbally, but also visually by exchanging a drawing of troop movements or other visual information to enhance the effectiveness of voice coordination. We are developing a secure voice system for the 21st century.

15. SUBJECT TERMS
Information hiding Vocoder
Tactical communication

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			George S. Kang
Unclassified	Unclassified	Unclassified	UL	17	**19b. TELEPHONE NUMBER** *(include area code)* (202) 767-2157

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

CONTENTS

HIDING INFORMATION UNDER SPEECH

1. INTRODUCTION

Speech communication is speedy and interactive. As a result, the telephone is an indispensable means of communication, particularly for the military. Until recently, secure voice has always been the military's fundamental element of command, control, and communications. With the fast convergence of all multiple forms of communications with computer networking, secure voice, by its very nature to human interaction, will continue to be the fundamental element of the coming enhanced collaborative communications environment.

Speech communication alone has disadvantages. Speech is a stream of sounds that is bound to a time continuum. We hear speech as it arrives in real time, and it disappears as fast as it arrives. Furthermore, our cognitive process for translating audio sounds to the meaning of the message is slow and imperfect. We often misunderstand (or mishear) messages, or we may even forget the vital part of the message, particularly voice communication that takes place in stressful situations. This is especially true for tactical communication, which often involves complicated troop maneuvers and/or fast-varying scenarios in chaotic environments.

The effectiveness of voice communication can be significantly improved if the listener is able to see key words, key phrases, or clearly drawn figures (all of which do not disappear as voice communication ends) simultaneously with the speech communication, in the same manner that a sports scoreboard and video display augment play-by-play announcing. While this may seem obvious to those accustomed to desktop computers with relatively high bandwidth networking, it is not generally a capability available to tactical secure voice users.

Tactical communication often involves complicated tactical maneuvers and/or fast-varying scenarios in imperfect input and listening environments. Tactical communication will become more effective if the communicator is able to see succinctly drawn charts and appropriate key words with speech. More importantly, the visual data may be sufficient to enable less spoken content, a quicker conclusion to the conversation, fewer follow-up questions, and less clarification. Less spoken content in turn ensures a greater opportunity for full attention for the entire duration of the message.

We are currently implementing a real-time demonstration model capable of transmitting visual aids simultaneously under continuous voice dialogue processed by the Variable Data Rate (VDR) voice encoder. This VDR voice encoder is currently under development by a Navy program office, SPAWAR PMW-160, for future Naval Voice over Internet Protocol (VoIP) applications. We expect that the communicability of naval tactical voice communication will be significantly improved by the use of the technology described in this report.

2. BACKGROUND DISCUSSIONS

2.1 The Term "Information Hiding" is a Misnomer

The term "information hiding" in the title of this project is widely used, but is somewhat of a misnomer for *our* applications. We are *not* interested in hiding information. Rather, we are interested in *transmitting*

Manuscript approved September 8, 2005.

data with speech without affecting the interoperability of voice terminals without this capability. An important requirement of our voice system is that the embedded information must be inaudible to listeners of the host speech, unless the embedded information is properly decoded and reassembled. We do not want to confuse this idea with steganography, whose goal is to make the embedded data completely undetectable. In addition, we must dismiss the idea of hiding data by using any technique that will require additional bandwidth. It is important to state clearly the constraints that must be satisfied for the specific applications in mind. Our applications are for simultaneously transmitting visual aids (such as key words, key phrases, or succinctly drawn figures) to improve the communicability of host speech without compromising interoperability with other voice terminals with the same voice algorithm but without the data-embedding capability.

2.2 Prior Information Hiding Efforts

While we are primarily interested in transmitting embedded information, there are enough similarities with hiding information that a review of past information hiding efforts is valuable. In 2001, there was the 4th International Workshop of Information Hiding [1]. In 2002, there was the 5th International Workshop on Information Hiding [2]. The review of these two documents and approximately 400 reference papers cited in these two workshop proceedings indicates that although information hiding research is considerably deep, prior information hiding efforts have been primarily for hiding information (often another image) under an *image*.

Scanning through Refs. 1 and 2 indicates that information hiding with speech has not been pursued in much depth. The plausible reasons might be: (1) speech is a signal with relatively small information content (typically, less than 100 kb/s, often much less, around 10 to 20 kb/s) whereas a one-page photograph contains as much as 4 Mb; therefore, an image has more room to hide data; and (2) speech steganography has not led to many money-making commercial businesses. For these two reasons, research in speech steganography has not flourished.

Recently, there have been two notable exceptions. The U.S. Air Force Research Laboratory (Rome, NY) has begun sponsoring the investigation of speech steganography [3]. Because of the difficulty in obtaining this reference, we quote the key objective of this project: "Development of an audio (*speech*) steganography demonstrator for ad hoc audio communication, for example IP-telephony, with a special focus on the real-time character of this medium." In addition, research at Polytechnic University in Brooklyn, NY, by Radhakrishnan, Shanmugasundaram, and Memon [4] has focused on data masking with the goal of masking "the secret message to make it appear like a normal multimedia object under typical statistical examinations." While this is important research, its goal of making random encrypted data "look" like audio is completely different than our goal of embedding data within speech.

As we discuss in Section 5, there are potentially many significant military applications of embedding data in speech that make tactical voice communication more effective. We feel that this data-embedding capability should be an important feature of any 21st century secure voice system. We have to demonstrate the usefulness of such a device.

Accordingly, we are expending considerable time to implement a user-friendly tactical voice system with an ability to transmit visual aids to the receiver. Our system is based on information hiding. The secondary data is embedded within the speech signal, and so, does not add any bandwidth to the overall communications channel and has no impact on the voice decoding of the receiver, allowing interoperation with voice terminals that do not have a data transmission capability.

3. TECHNICAL APPROACH

3.1 Interoperability Issue

Not every attempt to improve secure phones has resulted in a successful deployment because not all improved systems interoperate with other existing secure phones. To avoid a similar undesirable consequence, we imposed a few design constraints in order to achieve interoperability with similar secure phones that lack the data-embedding capabilities.

3.2 Approach to Embed Data under Speech

As discussed above, we use the information hiding approach to embed information because of the following reasons:

- No additional bandwidth is needed.
- No additional data rate is required.
- The embedded data is not audible or visible unless the embedded information is properly decoded and reassembled.

In our approach, each embedded data bit *replaces* the least significant bit (LSB) of an encoded speech parameter (therefore, no increased data rate is required). That is why secure voice with the data-embedding capability is transparent to the same secure voice system without data-embedding capability. Therefore, both secure voice systems are interoperable.

3.3 Difference between Data Hiding with Images and Data Hiding with Speech

Can we learn anything from data hiding with images that can be applied to data hiding with speech? Perhaps not too much, because the basic natures of speech and images are fundamentally different. Speech is a real-time phenomenon. Speech, as streaming data, is bound to the time continuum. Our ears receive it as it comes, instant by instant, and speech does not give us an opportunity of a posteriori examination. A photo that is ambiguous at first glance may be clarified by repeated viewing. There is no such possibility with speech. Therefore, hiding data under speech requires a different approach than hiding data within an image.

3.3.1 Exploitation of Human Auditory Perception Characteristics

Human auditory perception has certain peculiarities that must be exploited for hiding data effectively. For example, our ability to resolve tones decreases with the increase of frequency of the tone. Thus, it is more effective for hiding data in the higher frequency regions than in low frequencies. In addition, there is the phenomenon known as the “masking effect”; namely, the presence of a strong resonant frequency masks the perception of weaker frequencies near that resonant frequency. Thus, we can hide the data more effectively near the strong resonant frequencies. (In general, a preferred time to embed data is when speech is loud). Furthermore, the human ear is deaf to the phase of the speech signal, particularly to the stationary phase. This effect, which must be witnessed to be believed, can also be exploited.

3.3.2 Exploitation of Speech Waveform Cluster Characteristics

Clustering characteristics of the speech waveform in the frequency domain provide more possibilities for embedding data than in the time domain. The speech waveform in the frequency domain (Fig. 1) is

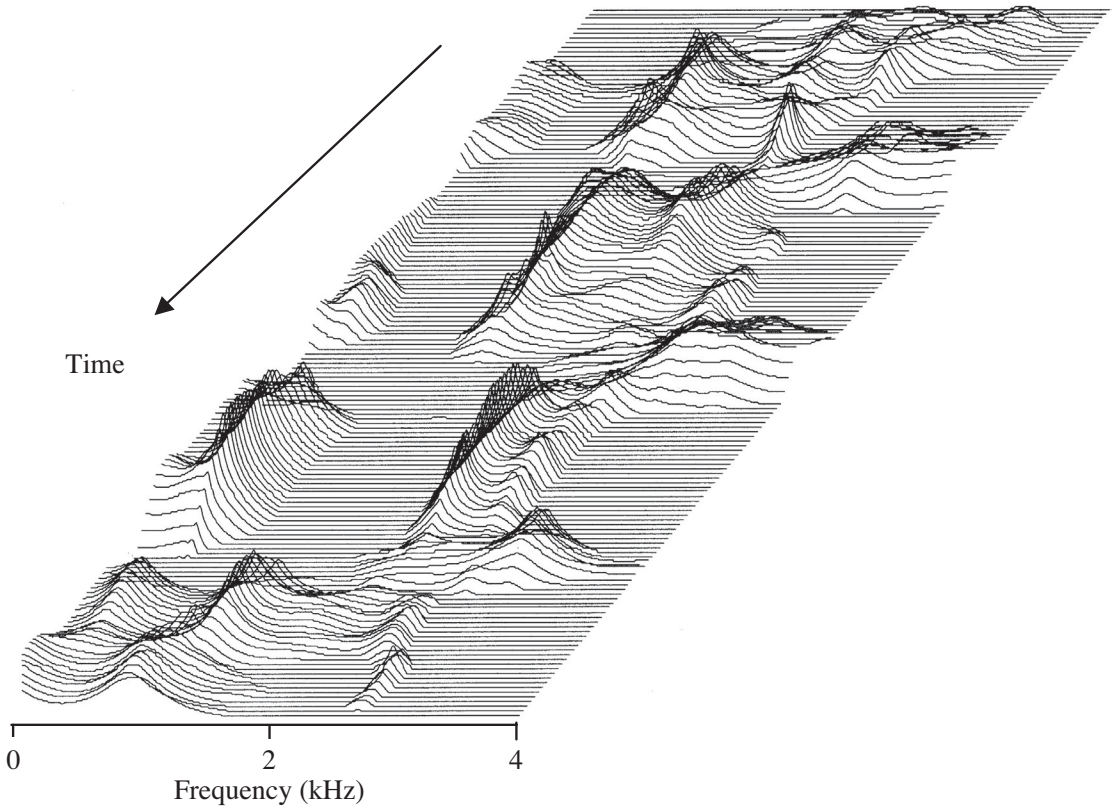

Fig. 1 — Running speech spectrum in time. Since vowels have three to four resonant frequencies, the speech spectrum has excellent cluster characteristics, which make embedding data more effective. As discussed in Section 4.3, we will embed data in the frequency domain of the host speech.

better clustered because vowels, a predominant component of speech, are made up of three to four resonant frequencies.

3.3.3 Embedded Data and Resultant Degradation of Host Speech

We need to establish a critical relationship between the amount of embedded data and speech quality degradation of the host speech. Since the host speech flows in real time at a certain data rate, the amount of embedded data is also best expressed in terms of "data rate" so that both quantities are expressed in the same unit. The most obvious example is when the host speech is encoded using Pulse Code Modulation (PCM) at 64 kb/s. Obviously, taking one bit from each 8-bit PCM sample at 8,000 samples/s gives us an embedded data of 8 kb/s. At this rate, we can embed another speech signal encoded at 8 kb/s under the 64-kb/s PCM host speech. This "bit-robbing" will have a negative effect on the host speech. In this simple case, further bit robbing will increase the degradation of the host speech quality in proportion to the number of bits that are "robbed." The degradation due to embedded data in more complex schemes will not be so clear-cut.

To establish such a relationship, we will rely on formalized methods for evaluating speech quality, such as the Diagnostic Acceptability Measure (DAM), or speech intelligibility, such as the Diagnostic Rhyme Test (DRT). These tests have been widely used for evaluating DoD secure voice systems. Both the DRT and DAM are listening tests conducted by the trained crew of a third party not affiliated with this research task. These tests provide highly detailed descriptions of speech quality or speech intelligibility.

3.3.4 On Reducing Audible Noise Generated by the Embedded Data

In our approach to embedding data, each bit of embedded data replaces the LSB of the encoded speech data. The effect is analogous to the fixed-point representation of floating point numbers. The quantization error is uniformly distributed over (0 to 99). Spectrally, the quantization error has a flat spectrum (it will make a "hiss" sound); therefore, such noise will be distinctively audible in the background of speech because vowels, a predominant part of speech, contain three to four resonant frequencies. Hence, reduction (or masking) of noise created by embedded data is a critical part of the technical approach.

What we do is to shape the noise spectrum in such a way that the noise will have a speech-like spectrum at the output so as to match the host speech. The concept is similar to the Dolby process for the analog audio recording process involving magnetic tapes. The "hiss" noise generated by the magnetic tape is spectrally flat; the idea is to boost high frequencies of the input and reduce high frequencies at the output using complementary filter characteristics. Therefore, the "hiss" is reduced with no effect on the input signal. However, in this process, high frequencies are attenuated at the output. To prevent this, we use more elaborate filtering, namely a combination of speech resonant frequency attenuation at the front end, and the speech resonant frequency amplification at the rear end. In other words, we whiten the speech spectrum at the front end and de-whiten the spectrum at the output (Figs. 2(a) and 2(b)).

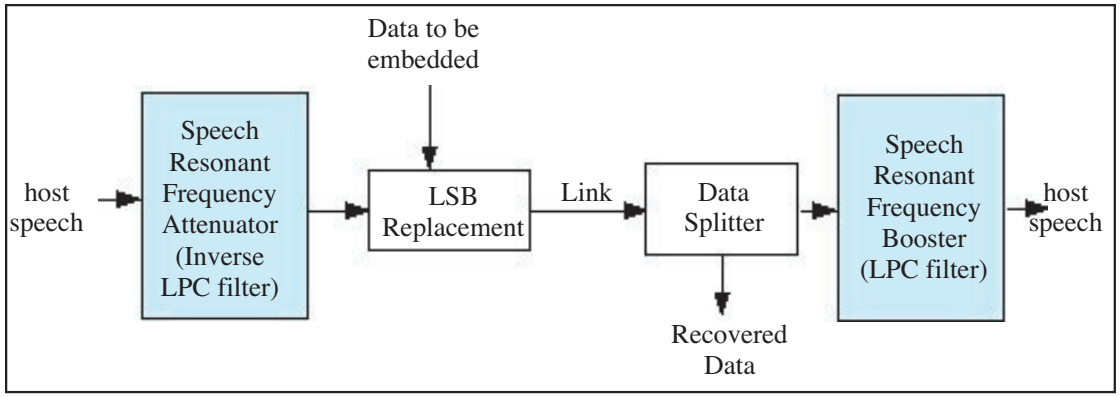

Fig. 2(a) — Spectral shaping process for the quantization error generated by embedding data in the LSB of host speech to reduce the audibility of noise at the output. The frequency responses of the spectral whitening filter and spectral de-whitening filter are shown in Fig. 2(b).

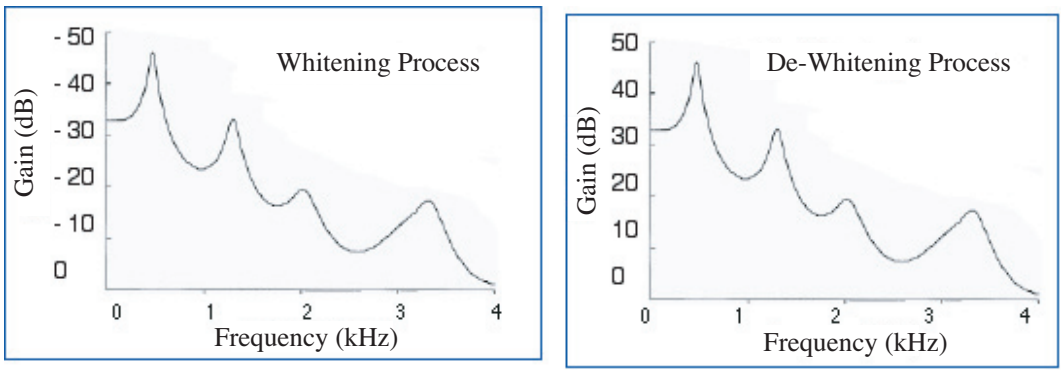

Fig. 2(b) — Frequency response of the Speech Resonant Frequency Booster. The frequency response is time-varying because filter weights are updated as often as 50 times/s. If a broadband signal is filtered by the spectral de-whitening filter, the resultant spectrum resembles the speech spectrum. Hence, resultant noise will be less audible in the presence of speech. Note that –dB of a gain is attenuation.

4. THREE FORMS OF HOST SPEECH

The host speech must be digital speech in order to make data embedding possible. Although there are seemingly infinite varieties of digital speech, we can group them into three generic categories: (1) wideband, (2) narrowband, and (3) mediumband speech encoder. The category of the host speech encoder directly affects the complexity of the overall system.

4.1 Wideband Speech Encoder

The most commonly used wideband speech encoder is PCM operating at 64 kb/s. It is widely used in the public switched telephone network. In PCM, analog-to-digital conversion is performed at 8,000 samples/s, and the speech amplitude is logarithmically represented into 8-bit samples. Thus, an 8-bit word is generated at every speech sampling time-interval of 125 μs. The data will be embedded in the LSB of this 8-bit word. Hence, the data rate of the embedded data is 8 kb/s.

The input-output amplitude transfer characteristics of 8-bit, μ-law PCM may be written in two different ways: (1) parametric representation and (2) the customary output in terms of input:

(1) Parametric representation:

$$x = \left(\frac{1}{2}\right)^n; \qquad y = \frac{8-n}{n}, \tag{1}$$

where $n = 0, 1, 2, 7$.

(2) Output as a function of input:

$$y \approx \frac{\ln(1+\mu x)}{\ln(1+\mu)}, \tag{2}$$

where $\mu = 255$ is the commonly used companding (compressing/expanding) factor. Equation (1) is a discrete representation, whereas Eq. (2) is a continuous (analog) representation, which is approximate, as indicated below.

$$\begin{array}{lll} x = 0 & y \text{ by Eq. (1)} = 0.875 & y \text{ by Eq. (2)} = 0.87503 \\ x = 0.25 & y \text{ by Eq. (1)} = 0.75 & y \text{ by Eq. (2)} = 0.752101 \end{array} \tag{3}$$

The amplitude quantization characteristics are shown in Fig. 3. The data embedding of the LSB is shown in Fig. 4.

4.2 Narrowband Speech Encoder

With a narrowband voice encoder (often called a *vocoder*), speech is not transmitted in terms of the time domain waveform. Rather, speech is transmitted in terms of speech parameters. These parameters are short-term averaged physical quantities such as the root-mean-square (RMS) energy value (indicating the loudness of speech), the pitch value, and most importantly, the resonant frequencies of the speech. In the past, various forms of narrowband vocoders were used. These vocoders were characterized by how the resonant speech frequencies were represented. For example, if the Fourier magnitudes over 16 channels are used, it is called a *channel vocoder*. If the all-pole digital filter (derived by linear prediction analysis) is used, it is called a *linear predictive coder* (LPC). It is significant to note that the narrowband vocoder parameters are

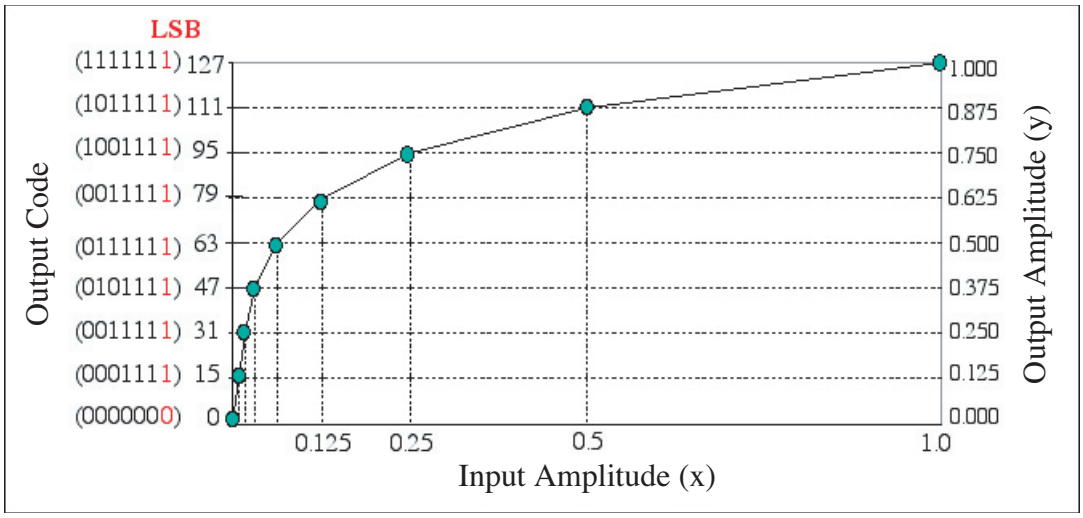

Fig. 3 — Input-output amplitude transfer characteristics of an 8-bit, μ-law PCM with μ = 255. As noted, if the peak amplitude is normalized to unity, the amplitude step corresponding to LSB is 0.125.

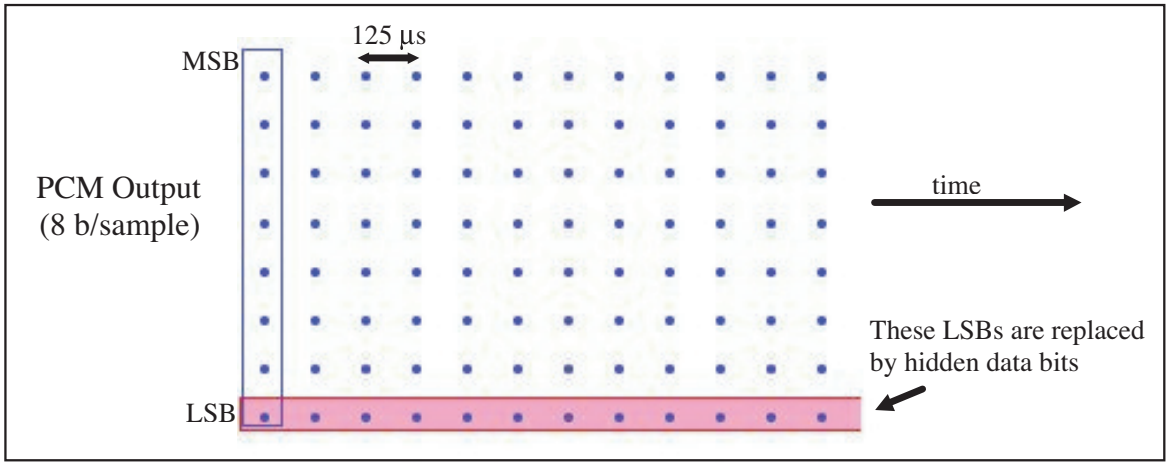

Fig. 4 — PCM generates 8 bits of speech amplitude every 125 μs, of which each LSB (indicated by the pink-colored box) is substituted with 1 bit of embedded data.

updated only once every 20 ms or so because these parameters do not change rapidly for normal conversational speech. Basically, all narrowband vocoders model the actual physical speech production mechanism shown in Fig. 5(a). The corresponding electrical model is shown in Fig. 5(b). Prior work [5] embedding data in narrowband speech is shown in Fig. 6.

4.3 Mediumband Vocoder

Mediumband vocoders are widely used to support tactical voice communication at data rates between 9.6 and 16 kb/s. We are most interested in mediumband vocoders because information hiding technology will benefit tactical voice communication at these rates the most. Although the most widely used mediumband vocoder is the Continuously Variable Slope Delta (CVSD) operating at 16 kb/s, we excluded CVSD from further discussion because its antiquated voice software/hardware has no expansion capabilities. A mediumband vocoder that has greater expansion potential is the Residual-Excited Linear Predictor (RELP). Our VDR vocoder developed for VoIP applications belongs to the RELP family. As discussed in the following paragraphs, we will embed data in the VDR vocoder.

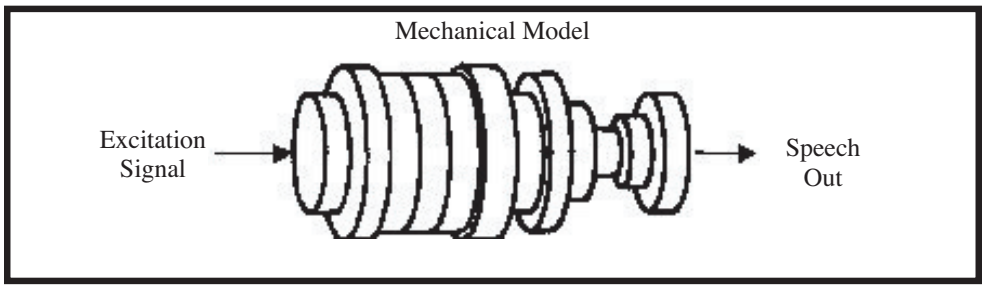

Fig. 5(a) — Mechanical model of the narrowband vocoder, in which the vocal tract is modeled as a concentric pipe that has a particular set of resonant frequencies.

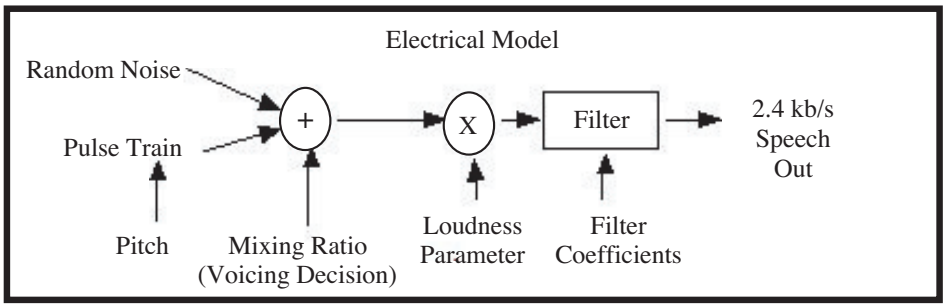

Fig. 5(b) — Electrical model of a narrowband vocoder. Data will be embedded in the LSBs of the speech parameters, mainly in the LSBs of the vocal tract filter coefficients—typically 10 coefficients for a total of 41 bits. They are updated as often as 44 times/s. As mentioned previously, the vocal tract filter has been modeled in many different ways.

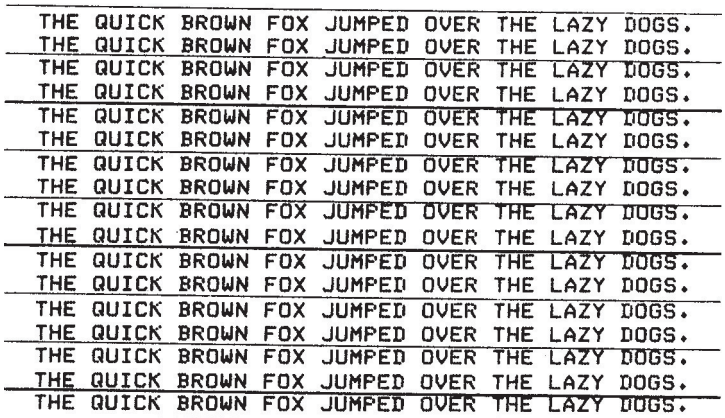

Fig. 6 — Actual text received at the rate of 80 b/s under a continuous AM broadcast encoded at 2400 b/s [5]. At this rate, we can transmit only a few key words, for example, enemy coordinates or the time of a planned event.

In PCM, the input speech waveform is directly quantized after a logarithmic amplitude transformation. In RELP, however, the speech signal is first spectrally flattened (i.e., whitened) by removing speech resonant frequencies and pitch harmonics. This spectrally whitened signal is called the *prediction residual*. After flattening, the prediction residual is then quantized. With PCM, speech samples must be quantized into as much as 6, 7, or 8 b/sample in order to have high-quality speech. In contrast, the RELP residual may be quantized into as coarse as 3 or 4 b/sample, yet still produce acceptable speech quality. This is because the de-whitening filter introduces speech resonant frequencies into the quantized residual, making the reconstruction of the original speech at the receiver more accurate. Accordingly, RELP is capable of generating high-quality speech at 16 kb/s, or even at 9.6 kb/s as demonstrated by Motorola's STU-III in 9.6-kb/s mode.

Placing the embedded data into the prediction residual makes good sense because the de-whitening filter introduces speech frequencies into the residual that mask the noise generated in the embedding process.

The prediction residual, $Y(z)$, in terms of the input speech, $X(z)$, is expressed by

$$Y(z) = \left[1 - \sum_{i=1}^{10} \alpha_i z^{-i}\right] X(z), \tag{4}$$

where α_i is the ith prediction coefficient derived from the input. The LSB of the prediction residual is replaced by each bit of embedded data.

The prediction residual with embedded data may be expressed as

$$Y'(z) = Y(z) + N(z), \tag{5}$$

where $Y(z)$ is the prediction residual of the input speech signal and $N(z)$ is the flickering LSB noises generated by the embedded data. $N(z)$ has a flat spectrum and uniformly distributed probability density function.

When the prediction residual with embedded data is passed through the de-whitening filter, which is identical to the LPC speech-synthesis filter, the filter introduces speech-resonant frequencies into the output. Thus, the output may be expressed as

$$E_o(z) = \frac{Y(z) + N(z)}{\left[1 - \sum_{i=1}^{10} \alpha_i z^{-i}\right]}, \tag{6}$$

which reduces to

$$E_o(z) = X(z) + \frac{N(z)}{\left[1 - \sum_{i=1}^{10} \alpha_i z^{-i}\right]}. \tag{7}$$

Equation (7) shows that the host speech output is a sum of the original input host speech and the spectrally shaped embedded data noise. The data noise is spectrally shaped to have the same spectrum as speech.

We are interested in the VDR vocoder [6] to embed data because it is currently being developed by a Navy program office, SPAWAR, PMW-160 for future VoIP applications. Figure 7 shows a flow diagram of VDR. Important characteristics of the VDR vocoder are the following:

- One voice processor and one crypto unit will provide universally interoperable wideband, medium-band, and narrowband speech.

- VDR generates 28 different instantaneous data rates in each frame of 22.5 ms. All of these data rates are in sync. Therefore, the data rate can be changed on the fly. VDR provides one optimum data rate based on (1) network traffic conditions, (2) the complexity of the speech waveform (vowels need more data than consonants), and (3) frequency-dependent auditory perception characteristics.

- VDR encodes the residual in the frequency domain because human auditory perception characteristics are frequency dependent. Likewise, data bits are embedded in the frequency domain. VDR transmits 96 spectral components (real and imaginary parts, or amplitude and phase components). Thus, each spectral component is spaced 41.667 Hz apart.

- Because of the variable nature of the data rate, each spectral component gets a different resolution in each frame. Three bits or more are dependent on the parameter that indicates the complexity of speech waveform (which is the maximum value of the prediction residual sample in each frame). Table 1 shows how the spectral resolution of the excitation signal varies with the waveform complexity index.

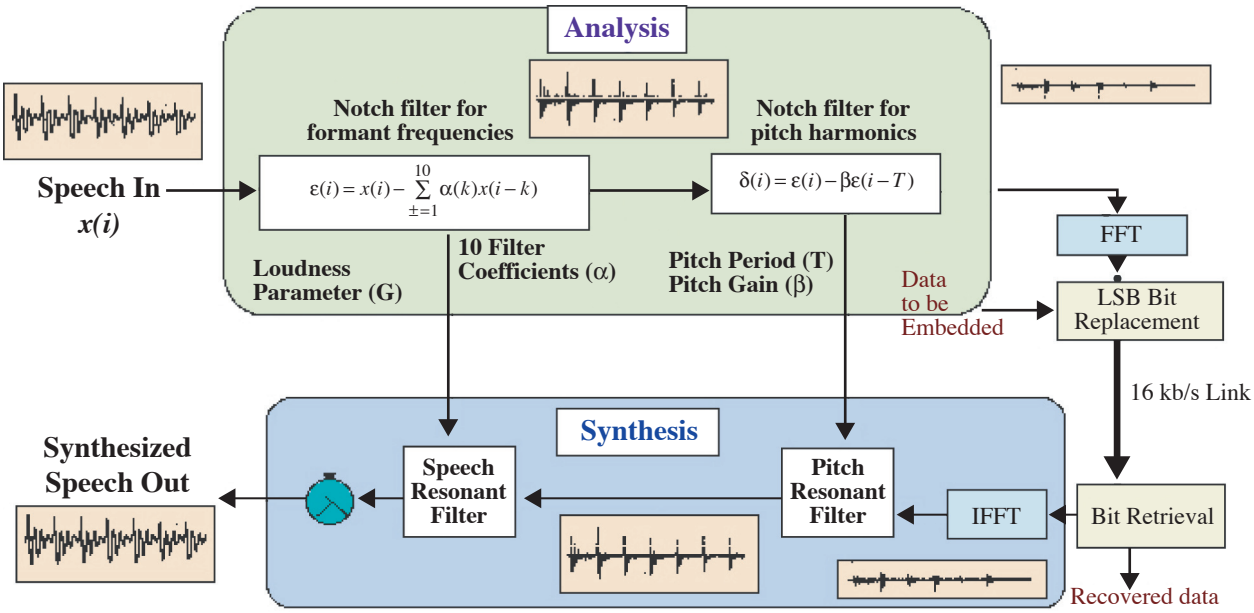

Fig. 7 — Flow diagram of the VDR vocoder. Data will be embedded in the LSB of the prediction residual of the host speech in the frequency domain. The spectrum of flickering LSB noise is automatically shaped by the LPC filter. Hence, embedded data noise is substantially less audible in this system. Since the VDR vocoder is currently under development by SPAWAR, PMW-160, it is an ideal vehicle to introduce the data embedding capability. The loudness parameter (G), determines the waveform complexity index shown in Table 1.

Table 1 — Index that Indicates the Speech Waveform Complexity and Spectral Resolution of the Excitation Signal

Spectral Magnitude of Pitch-Filtered Prediction Residual	Waveform Complexity Index (A)		Spectral Resolution of Excitation Signal		
			0 - 1 kHz	1 - 2 kHz	2 - 4 kHz
359 - 511	Complex	1	9 bits	8 bits	7 bits
148 - 358	↑	2	8	7	6
73 - 147		3	7	6	5
36 - 72		4	6	5	4
18 - 35		5	5	4	3
5 - 17	↓	6	4	3	–*
0 - 4	Simple	7	3	–*	–*

*These spectral components are not transmitted because their contribution to synthesized speech is insignificant. At the receiver, random amplitude and phase spectral components are used for these.

5. POTENTIAL APPLICATIONS

5.1. Transmission of a Tactical Assault Plan between Two Commanders

One useful application of embedding data under speech would be the simultaneous transmission of tactical coordination discussions while using hand-drawn tactical assault plans overtop a map (e.g., hand-sketches indicating the desired movements of troops, similar to John Madden using arrows to explain a football game in progress) as shown in Fig. 8.

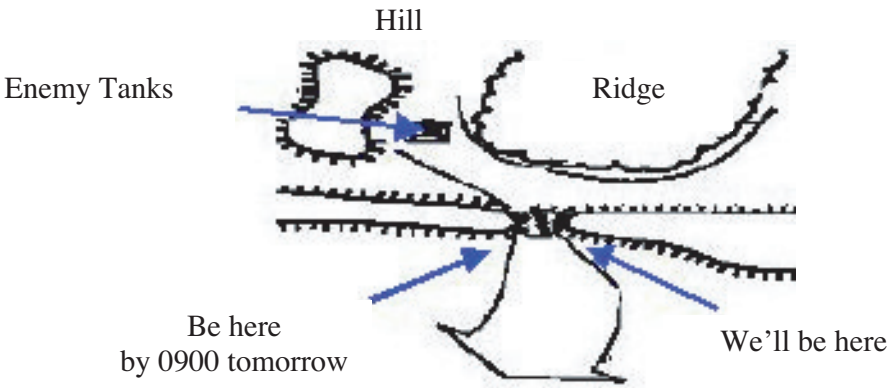

Fig. 8 — Tactical plan simultaneously transmitted while talking over tactical secure phone between two commanders connected by a tactical voice link. Availability of such graphics makes the tactical plan less ambiguous. Furthermore, such a graphics supplement makes it possible to reexamine the plan long after the communication is over. All written-in information is spoken over a secure phone, and all hand-written information (arrows, etc.) is transmitted as quickly as it is drawn.

5.2 Transmission of a Reconnaissance Observer Report

As an old saying goes, "a picture is worth a thousand words." Depending on the observer's verbal skill, sending a picture may take less time than explaining the same object verbally (Fig. 9). A photograph that is ambiguous at first glance may be clarified by repeated viewing. There is no such possibility with real-time speech.

Fig. 9 — A photograph of an object sighted by a reconnaissance observer. How can anyone explain this object over the phone? Even a low-quality photograph would be a preferred visual aid. The data-under-voice technology would be highly beneficial in similar applications encountered in other tactical situations. By the way, the object is an old Soviet bomber.

5.3 Transmission of a Naval Offshore Bombardment Observation Report

Naval offshore bombardment is often monitored by low-flying aircraft or helicopters over the target area. The observer reports back on the accuracy of the shelling. Reporting is difficult unless the hit coordinates can be indicated on a map (Fig. 10). The actual bombardment coordinates and predetermined target location (indicated in Fig. 10 by X on the map) can be sent along with any verbal dialogue customarily exchanged for the mission.

Fig. 10 — Example of an offshore bombardment observation report. In addition to a voice report, it would be preferred to send a map or hit coordinates similar to Fig. 8, which will enhance the offshore bombardment operation.

5.4 Transmission of the Speaker's Fingerprint for Information Assurance Purpose

In the early 2000s, biometrics became a major DoD thrust. On July 13, 2000, the President of the United States directed the DoD to create a program office to develop and promulgate biometrics technologies to improve information system security [7]. Accordingly, DoD established the Biometrics Management Office (BMO) with the Army as its executive agent.

Although there are many forms of biometrics for identifying a person, biometrics based on fingerprinting is perhaps the most mature technology. Fingerprint-based biometrics is reliable because fingerprints are basically time-invariant (unlike a person's voice). Access to secure voice terminals and communications links can be controlled by transmitting the user's fingerprint, such as the one shown in Fig. 11(a), along with their speech. Embedding this fingerprint data within existing vocoders allows for access control, yet keeps backward compatibility with existing secure voice equipment.

(a) (b)

Fig. 11(a) — Transmission of the speaker's fingerprint for information assurance. A specific biometrics application is shown in Fig. 11(b), in which a soldier wishes to be picked up by a friendly helicopter that wants to know the identity of the soldier. Dr. Jim Davies, an erstwhile director of Emerging Technologies at SPAWAR, PMW-161 stated, "If this type of tactical secure voice terminal were available, it would be a hot-selling item among services." We intend to develop such a voice terminal.

5.5 Transmission of Medical Photos

When a front-line physician needs to consult with an expert behind the line (for example, experts at Walter Reed Hospital in Washington, DC) over a voice link, the simultaneous transmission of medical photos (e.g., X-ray or MRI photos) would be a benefit to making a remote diagnosis (Fig. 12).

Fig. 12 — MRI photo in 28 kb. This type of photo may be transmitted under speech during a voice conversation in progress.

6. CONCLUSIONS

Over the past 50 years, the DoD has developed well over 100 different secure telephones. Unfortunately, they provided nothing more to us than did the original telephone known to us since its invention in 1883; that is, the secure phone is a device capable of transporting only sounds (the human voice and surrounding sounds).

Recently, the art of presentation and communicability have been studied in depth. According to an expert in this field, "simple but accurate charts are essential to make a presentation more effective." In other words, secure phones of the 21st century must be able to transmit visual aids with speech. Then, tactical voice co-

ordination will become more effective, and warfighting capability will improve. This report presents a new way of embedding data under tactical voice terminals operating at a data-rate range of 9.6 to 16 kb/s. The strengths of our method are the following:

- It does not require additional bandwidth beyond what is required to transmit speech alone.
- The voice terminal with the data-embedding capability will still interoperate with a similar voice terminal without the data-embedding capability.
- The quality of the host speech will not be degraded by the presence of the embedded data.

7. ACKNOWLEDGMENTS

This project was entirely sponsored by the Office of Naval Research as part of an NRL Base Program. The authors express their appreciation to Mike Weber and Bill Kordela of the Navy Information Security Office (SPAWAR PMW-161), who earlier sponsored NRL to develop the VDR vocoder that became the vehicle for embedding data under speech for the present research.

REFERENCES

1. I.S. Moskowitz, ed., *Proceedings: 4th International Workshop, Information Hiding (IH) 2001, April 25-27, 2001* (Springer-Verlag, New York, 2002).

2. F.A.P. Petitcolas, ed., *Proceedings: 5th International Workshop, Information Hiding (IH) 2002, October 7-9, 2002* (Springer-Verlag, New York, 2003).

3. J. Dittmann, Otto-von-Guericke University of Magdeburg, Document sent to U.S. Air Force Materiel Command—Rome Research Site, AFRL/Information Directorate, 26 Electronic Parkway, Rome, New York, December 2002.

4. R. Radhakrishnan, K. Shanmugasundaram, and N. Memon, "Data Masking: A Secure-Covert Channel Paradigm," IEEE Workshop on Multimedia Signal Processing, December 9-11, 2002, St. Thomas, Virgin Islands, USA.

5. G.S. Kang, "Narrowband Integrated Voice Data System Based on the 2400-B/S LPC," NRL Report 8942, December 1985.

6. G.S. Kang, "Variable-Data-Rate Voice Encoder for Voice Over Internet Protocol (VoIP)," NRL/FR-MM/5550--01-10,016, December 2001.

7. U.S. Public Law 246. 106th Cong., 2nd Sess., 13 July 2000, Military Construction Appropriations Act of 2001. Http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=106_cong_public_laws&docid=f:publ246.106.pdf